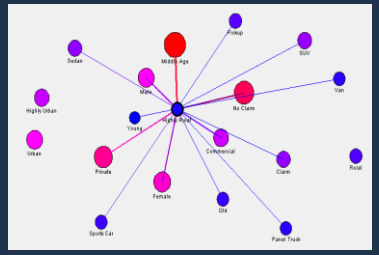
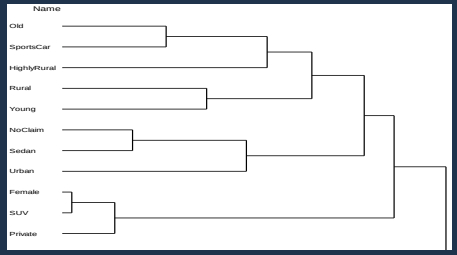




Data Science in Action #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Gerhard Svolba
Data Scientist, SAS Austria

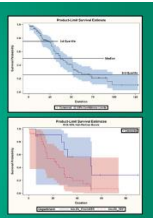
sas
THE POWER TO KNOW.

Data Science Applications and Case Studies

Data Science
in Action: #1

Performing Headcount Survival Analysis for Employee Retention

Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?

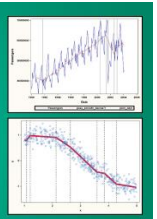


Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

Data Science
in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

Can events and changes in the course over time be automatically detected?

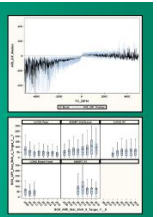


Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

Data Science
in Action: #3

Explaining Forecast Errors and Deviations

Do the demand planners really improve forecast accuracy with their manual overwrites?

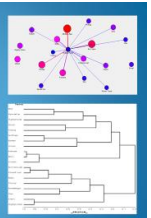


Linear Regression
Quantile Regression
Descriptive Statistics

Data Science
in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?

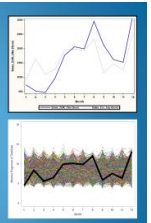


Unsupervised machine learning methods:
association analysis
variable clustering

Data Science
in Action: #5

Checking the Alignment with Predefined Pattern

Which customers show a behavior that is far from what you expected?

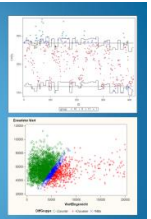


Chi2 independency test
Benford's law
Time Series Similarity

Data Science
in Action: #6

Proving a reference value that considers all available co-information

Can analytics help me to reduce the "Yes, but ... " sentences in my business discussions?

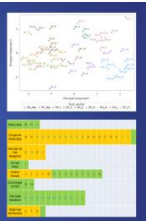


Linear Regression
Decision Trees
Time Series Analysis

Data Science
in Action: #7

Topic Search Documents and Clustering

Can I automatically find clusters of documents with similar content?

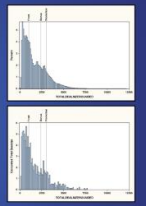


Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

Data Science
in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?

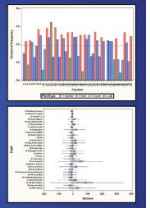


Monte Carlo Simulations
Mathematical Programming

Data Science
in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game

How can you simulate complex environments to get insight in the most frequent processes?

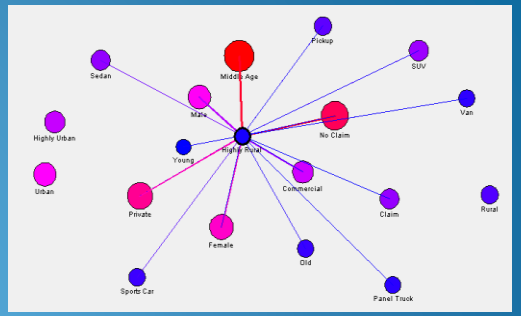


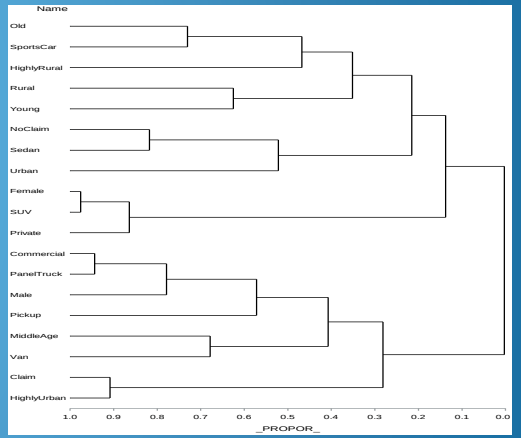
Monte Carlo Simulations

Data Science in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?





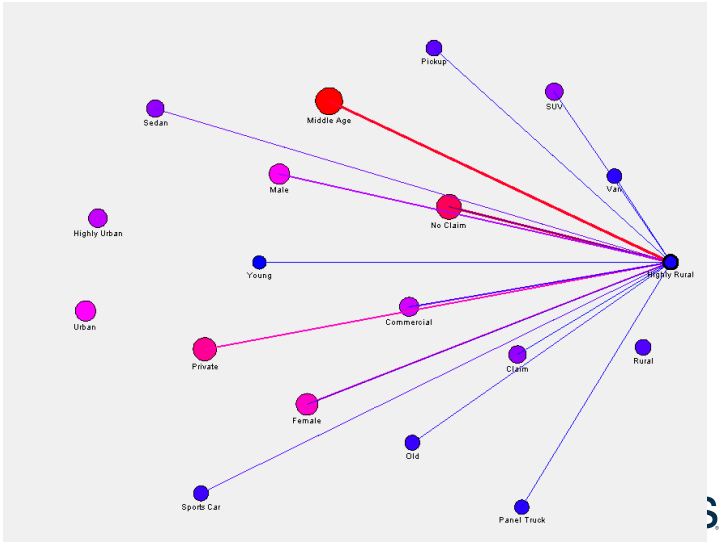
Unsupervised machine learning methods:
association analysis
variable clustering

Make your Data Talk to You!

- Data from Car Insurance with 6 properties per customer

Variable	Feature
AGE	YOUNG, MIDLIFE, OLD
GENDER	MALE, FEMALE
DENSITY	HIGHLY URBAN, URBAN, HIGHLY RURAL, RURAL
CAR_TYPE	VAN, SPORTS CAR, SUV, SEDAN, PICK UP
CAR_USAGE	PRIVATE, COMMERICIAL
CLM_FLAG	CLAIM, NO CLAIM

- Use unsupervised machine learning (association analysis) to uncover relationships between different properties.



Background and Usage Instructions

- You do not have to ask every question separately.
  But you can let your data talk.
- Can also be used in combination:
  - Get insight into your data, check data quality
    → Unsupervised ML
  - Create the predictive model → Supervised ML
- Important for the interpretation:
  This is explanatory data analysis!
  - It is not a controlled experiment, where a dedicated questions has been predefined.
  - It does not test hypotheses!
- Yes, you will also find obvious rules.
  - It is in the nature of things.

Try it! Transpose your data in a way you usually would not do it.

One-Row-Per-Subject

POLICYNO	CLM_FLAG	CAR_USE	CAR_TYPE	AGE	GENDER	DENSITY
160	No	Private	Sedan	60	M	Highly Urban
24836	No	Commercial	Sedan	43	M	Highly Urban
28046	No	Private	Van	48	M	Urban
28960	No	Private	SUV	35	F	Highly Urban
40933	No	Private	Sedan	51	M	Highly Urban
55277	No	Private	SUV	50	F	Urban
63212	Yes	Commercial	Sports Car	34	F	Highly Urban
69651	No	Private	SUV	54	F	Highly Urban
88070	Yes	Private	Sedan	40	M	Urban
93553	No	Commercial	SUV	44	F	Rural
127444	Yes	Commercial	Van	37	M	Highly Urban
141509	Yes	Private	SUV	34	F	Highly Urban
145326	No	Commercial	Van	50	M	Rural
146809	Yes	Private	Sports Car	53	F	Urban
148250	No	Private	Sedan	43	F	Rural
157851	No	Commercial	Van	55	M	Urban

Multple-Row-Per-Subject Key-Value Table

POLICYNO	Feature
160	Highly Urban
160	No Claim
160	Sedan
160	Private
160	Male
160	Old
24836	Highly Urban
24836	No Claim
24836	Sedan
24836	Commercial
24836	Male
24836	Middle Age

The 1-0-1 of Interpreting the Results of Association Analyses

- **Rule**: Milk → Cake (LHS → RHS)
- **Confidence**: Customers, who by milk (LHS) buy in 23% of the cases also cake (RHS)
- **Support**: relative frequency of „ milk + cake“ combinations in all baskets. eg.: 4.67 %
- **Lift**: 2.3 = multiplicative factor that the rule „milk → cake“ appears more frequent, than under the assumpation of indepency.
- Important: both extremes of the lift are of interest here: Lift >> 1 und Lift << 1.

Looking at the results table

index	RULE	_LHAND	_RHAND	COUNT	SUPPORT	EXP_CONF	CONF	LIFT	PVALUE
1	Commercial ==> Panel Truck	Commercial	Panel Truck	853.00	8.28	8.28	22.51	2.72	0.0000
2	Panel Truck ==> Commercial	Panel Truck	Commercial	853.00	8.28	36.78	100.00	2.72	0.0000
3	Young ==> Claim	Young	Claim	78.00	0.76	26.66	65.00	2.44	0.0000
4	Claim ==> Young	Claim	Young	78.00	0.76	1.16	2.84	2.44	0.0000
5	Panel Truck ==> Male	Panel Truck	Male	813.00	7.89	46.18	95.31	2.06	0.0000
6	Male ==> Panel Truck	Male	Panel Truck	813.00	7.89	8.28	17.09	2.06	0.0000
7	Van ==> Male	Van	Male	804.00	7.80	46.18	87.30	1.89	0.0000
8	Male ==> Van	Male	Van	804.00	7.80	8.94	16.90	1.89	0.0000
9	Sports Car ==> Female	Sports Car	Female	1149.0	11.15	53.82	97.46	1.81	0.0000
10	Female ==> Sports Car	Female	Sports Car	1149.0	11.15	11.44	20.72	1.81	0.0000

Men Do Not Drive Sports Cars?

Rule 278 shows that sports cars are only driven by men in 2.54% of the cases, whereas this was expected in around 46% of the cases.

index	RULE	_LHAND	_RHAND	COUNT	SUPPORT	EXP_CONF	CONF	LIFT
267	Commercial ==> Sports Car	Commercial	Sports Car	200.00	1.94	11.44	5.28	0.46
268	Rural ==> Claim	Rural	Claim	102.00	0.99	26.66	6.52	0.24
269	Claim ==> Rural	Claim	Rural	102.00	0.99	15.18	3.71	0.24
270	Young ==> Highly Urban	Young	Highly Urban	10.00	0.10	34.93	8.33	0.24
271	Highly Rural ==> Claim	Highly Rural	Claim	32.00	0.31	26.66	6.30	0.24
272	Claim ==> Highly Rural	Claim	Highly Rural	32.00	0.31	4.93	1.17	0.24
273	Van ==> Female	Van	Female	117.00	1.14	53.82	12.70	0.24
274	Female ==> Van	Female	Van	117.00	1.14	8.94	2.11	0.24
275	Panel Truck ==> Female	Panel Truck	Female	40.00	0.39	53.82	4.69	0.09
276	Male ==> SUV	Male	SUV	99.00	0.96	27.98	2.08	0.07
277	SUV ==> Male	SUV	Male	99.00	0.96	46.18	3.43	0.07
278	Sports Car ==> Male	Sports Car	Male	30.00	0.29	46.18	2.54	0.06

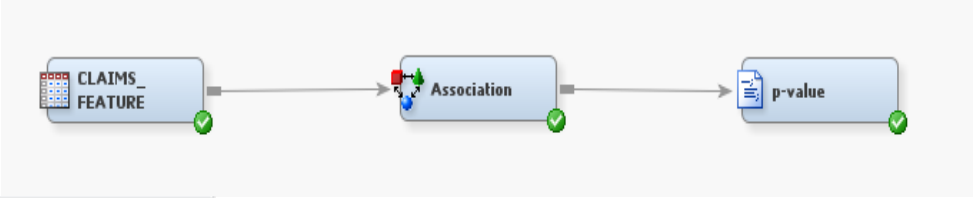
- This might indicate a situation that for the customer base, sports cars are really predominantly driven by women.
- It could be a trigger to an investigation of the quality status of your data.
- A business interpretation could be that in a family, the sports car is the 2nd or 3rd car that is registered in the wife's name for financial reasons.
- The competitor is offering a policy to men for a much more attractive price.

Association-Analysis with SAS

Running Association Analysis in SAS

- SAS Enterprise Miner
  - Association Node
  - PROC ASSOC
- SAS Viya (VDMML)
  - PROC MBANALYSIS
  - MBANALYSIS action

Use the Associations Node in SAS Enterpirse Miner



ID	claimsfeature
Name	CLAIMS_FEATURE
Variables	
Decisions	
Role	Transaction

Variables - CLAIMS_FEATURE

(Ohne) not Ist gleich

Columns: Label

Name	Role	Level
Feature	**Target**	**Nominal**
POLICYNO	**ID**	**Interval**

Association	
Maximum Items	2
Minimum Confidence Level	1
Support Type	Count
Support Count	1
Support Percentage	5.0
Sequence	
Rules	
Number to Keep	10000
Sort Criterion	Default
Number to Transpose	5000
Export Rule by ID	No
Recommendation	No

Use the MBANALYSIS procedure or the MBANALYSIS CAS-Action in SAS Viya

```
proc mbanalysis
  data=casdata.claims_feature
   items    = 2
   support  = 1
   conf     = 1
   lift     = 0.001 ;
 customer policyno;
 target   feature;
 output outrule=
    casdata.claims_rules;
run;
```

```
Proc CAS;
 ruleMining.mbanalysis /
   table={name='CLAIMS_FEATURE',
          caslib='casdata'},
   supmin=1,
   idVariable='POLICYNO',
   tgtVariable='Feature',
   hierarchy={},
   outrule={name='CLAIMS_RULES',
            caslib='casdata'
            replace=true},
   conf=1,
   items=2,
   lift=0.001;
Quit;
```

Use the Variable-Clustering Method

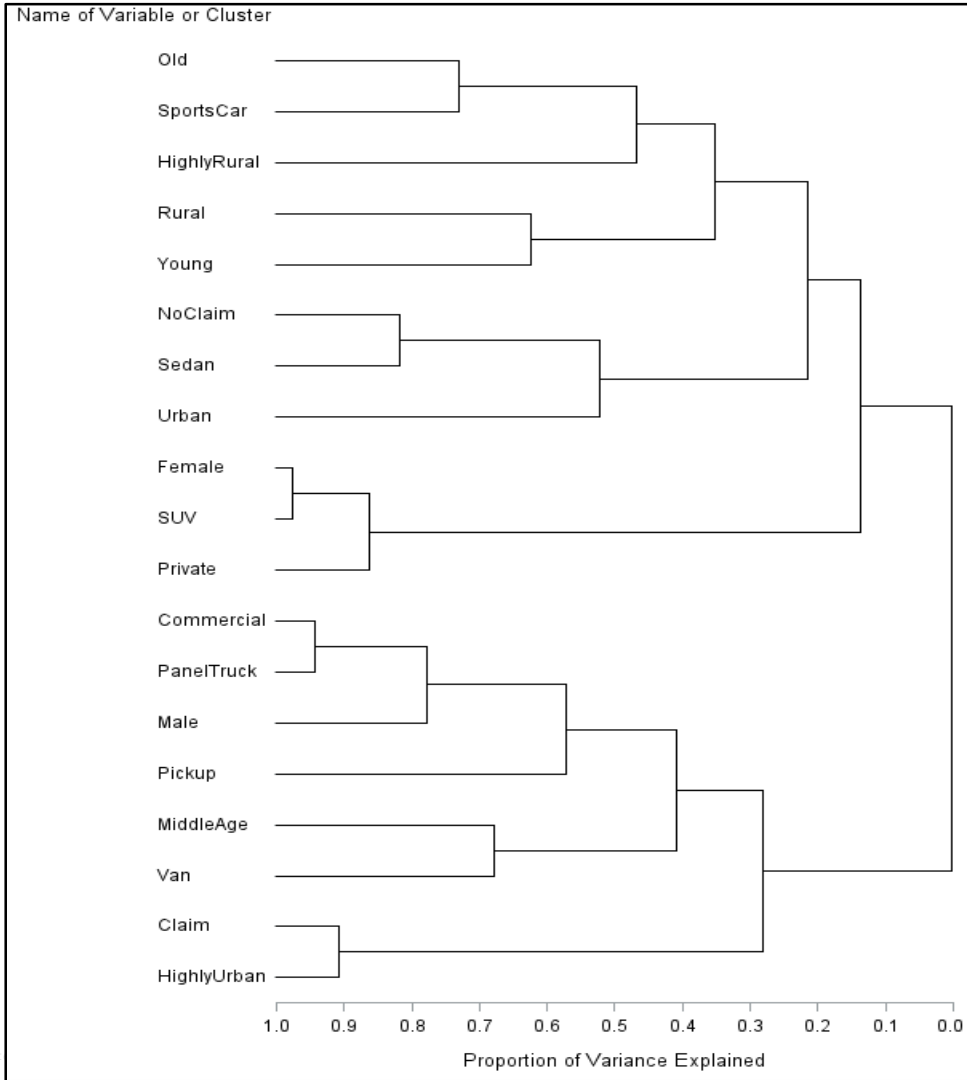
PROC Varclus: Data Structure and Code

POLICYNO	HighlyUrban	NoClaim	Sedan	Private	Male	Old	Commercial	MiddleAge	Urban	Van	SUV	Female	Claim
160	1	1	1	1	1	1	0	0	0	0	0	0	0
24836	1	1	1	0	1	0	1	1	0	0	0	0	0
28046	0	1	0	1	1	0	0	1	1	1	0	0	0
28960	1	1	0	1	0	0	0	1	0	0	1	1	0
40933	1	1	1	1	1	0	0	1	0	0	0	0	0
55277	0	1	0	1	0	0	0	1	1	0	1	1	0
63212	1	0	0	0	0	0	1	1	0	0	0	1	1
69651	1	1	0	1	0	0	0	1	0	0	1	1	0
88070	0	0	1	1	1	0	0	1	1	0	0	0	1
93553	0	1	0	0	0	0	1	1	0	0	1	1	0

```
proc varclus data=claims_1row
             outtree=varclus_tree1
             centroid;
 var Claim Commercial Female HighlyRural HighlyUrban
     Male MiddleAge NoClaim Old PanelTruck Pickup
     Private Rural SUV Sedan SportsCar Urban Van Young;
run;
```

Results of Variable Clustering

5 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	Claim	0.6620	0.3765	0.5421
	HighlyUrban	0.6620	0.1311	0.3890
Cluster 2	HighlyRural	0.1761	0.0279	0.8475
	MiddleAge	0.2189	0.4460	1.4100
	NoClaim	0.3765	0.6620	1.8447
	Rural	0.2074	0.0954	0.8762
	Sedan	0.2167	0.0455	0.8206
Cluster 3	Commercial	0.5344	0.4475	0.8427
	Male	0.5320	0.6487	1.3322
	PanelTruck	0.2696	0.1513	0.8607
	Pickup	0.1690	0.1017	0.9250
	Van	0.1756	0.0731	0.8894
Cluster 4	Old	0.2741	0.1874	0.8932
	SportsCar	0.2732	0.0908	0.7994
	Urban	0.2759	0.2346	0.9460
	Young	0.2546	0.0266	0.7657
Cluster 5	Female	0.6487	0.5320	0.7506
	Private	0.4475	0.5344	1.1867
	SUV	0.6098	0.2582	0.5261

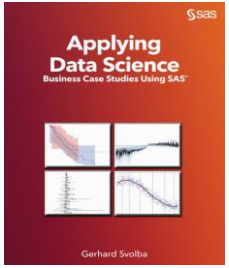


Conclusion

- Unsupervised machine learning techniques provide insight into your data.
- Well-known relationships, anomalies, interesting findings
- The data talk to you! You do not have to ask them feature by feature.
- Note that this is explanatory data analysis!

Analytics and Data Science is there to help you!

- Get a clearer, more objective picture of your data and your analysis subjects
- Get explicit results instead of searching the needle in the haystack
- Make your data talk to you!
- Receive findings automatically instead of manually
- Do it again! – treat models as an asset and repeat your analysis

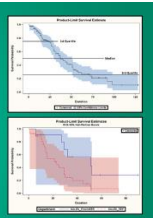


Data Science Applications and Case Studies

Data Science
in Action: #1

Performing Headcount Survival Analysis for Employee Retention

Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?

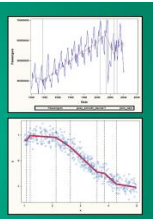


Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

Data Science
in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

Can events and changes in the course over time be automatically detected?

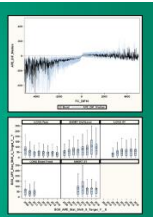


Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

Data Science
in Action: #3

Explaining Forecast Errors and Deviations

Do the demand planners really improve forecast accuracy with their manual overwrites?

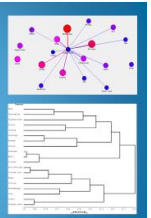


Linear Regression
Quantile Regression
Descriptive Statistics

Data Science
in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?

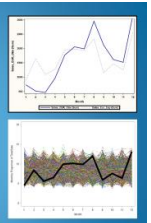


Unsupervised machine learning methods:
association analysis
variable clustering

Data Science
in Action: #5

Checking the Alignment with Predefined Pattern

Which customers show a behavior that is far from what you expected?

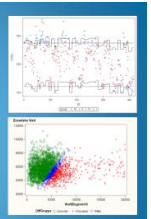


Chi2 independency test
Benford's law
Time Series Similarity

Data Science
in Action: #6

Proving a reference value that considers all available co-information

Can analytics help me to reduce the "Yes, but ... " sentences in my business discussions?

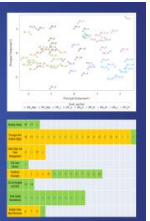


Linear Regression
Decision Trees
Time Series Analysis

Data Science
in Action: #7

Topic Search Documents and Clustering

Can I automatically find clusters of documents with similar content?

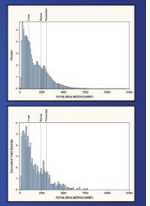


Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

Data Science
in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?

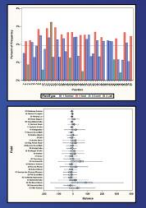


Monte Carlo Simulations
Mathematical Programming

Data Science
in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game

How can you simulate complex environments to get insight in the most frequent processes?



Monte Carlo Simulations

Get access to more content:

SAS DACH @Youtube:	https://www.youtube.com/user/SASsoftwareGermany
Blogs on LinkedIn:	https://www.linkedin.com/in/gerhardsvolba/
Twitter:	https://twitter.com/gsvolba
Content on Github:	https://github.com/gerhard1050
Books @SAS-Press:	https://support.sas.com/svolba